Transformer-based Argument Mining for Healthcare Applications

Tobias Mayer, Elena Cabrio, Serena Villata

Transformer-based Argument Mining for Healthcare Applications

Tobias Mayer and **Elena Cabrio** and **Serena Villata**[1]

Abstract. Argument(ation) Mining (AM) typically aims at identifying argumentative components in text and predicting the relations among them. Evidence-based decision making in the healthcare domain targets at supporting clinicians in their deliberation process to establish the best course of action for the case under evaluation. Although the reasoning stage of this kind of frameworks received considerable attention, little effort has been devoted to the mining stage. We extended an existing dataset by annotating 500 abstracts of Randomized Controlled Trials (RCT) from the MEDLINE database, leading to a dataset of 4198 argument components and 2601 argument relations on different diseases (i.e., *neoplasm*, *glaucoma*, *hepatitis*, *diabetes*, *hypertension*). We propose a complete argument mining pipeline for RCTs, classifying argument components as *evidence* and *claims*, and predicting the relation, i.e., *attack* or *support*, holding between those argument components. We experiment with deep bidirectional transformers in combination with different neural architectures (i.e., LSTM, GRU and CRF) and obtain a macro F1-score of .87 for component detection and .68 for relation prediction, outperforming current state-of-the-art end-to-end AM systems.

1 Introduction

In the healthcare domain, there is an increasing interest in the development of *intelligent* systems able to support and ease clinicians' everyday activities. These systems apply to clinical trials, clinical guidelines, and electronic health records, and their solutions range from the automated detection of PICO[2] elements [19] in health records to evidence-based reasoning for decision making [18, 8, 24, 35]. These applications highlight the need of clinicians to be supplied with frameworks able to extract, from the huge quantity of data available for the different diseases and treatments, the exact information they necessitate and to present this information in a structured way, easy to be (possibly semi-automatically) analyzed. Argument(ation) Mining (AM) [30, 22, 7] deals with finding argumentative structures in text. Standard tasks in AM consist in the detection of argument components (i.e., *evidence* and *claims*), and the prediction of the relations (i.e., *attack* and *support*) holding among them. Given its aptness to automatically detect in text those argumentative structures that are at the basis of evidence-based reasoning applications, AM represents a potential valuable contribution in the healthcare domain.

However, despite its natural employment in healthcare applications, only few approaches have applied AM methods to this kind of text [14, 25, 26], and their contribution is limited to the detection of argument components, disregarding the more complex phase of predicting the relations among them. In addition, no huge annotated dataset for AM is available for the healthcare domain. In this paper, we cover this gap, and we answer the following research question: *how to define a complete AM pipeline for clinical trials?* To answer this question, we propose a deep bidirectional transformer approach combined with different neural networks to address the AM tasks of component detection and relation prediction in Randomized Controlled Trials, and we evaluate this approach on a new huge corpus of 659 abstracts from the MEDLINE database.

More precisely, the contributions of this paper are as follows:

1. We build a new dataset from the MEDLINE database, consisting of 4198 argument components and 2601 argument relations on five different diseases (*neoplasm*, *glaucoma*, *hepatitis*, *diabetes*, *hypertension*)[3];
2. We present a complete AM pipeline for clinical trials relying on deep bidirectional transformers combined with different neural networks, i.e., Long Short-Term Memory (LSTM) networks, Gated Recurrent Unit (GRU) networks, and Conditional Random Fields (CRFs)[4];
3. Our extensive evaluation of various AM architectures (e.g., for persuasive essays) reveals that current approaches are unable to adequately address the challenges raised by medical text and we show that transformer-based approaches outperform these AM pipelines as well as standard baselines.

In the following, Section 2 compares the proposed approach with related work. We then describe the corpus we built, the methods we employed and the experimental setting. Finally, we report the obtained results and we address an error analysis before concluding.

2 Related work

One of the latest advances in artificial argumentation [2] is the so-called *Argument(ation) Mining* [30, 22, 7]. Argument mining consists of two standard tasks: *(i)* the identification of arguments within the text, that may be further split in the detection of argument components (e.g., claims, evidence) and the identification of their textual boundaries. Different methods have been used for this task (e.g., Support Vector Machines (SVMs), Naïve Bayes classifiers, and Neural Networks (NNs)); *(ii)* the prediction of the relations holding between the arguments identified in the first stage. They are used to build the

[1] Université Côte d'Azur, CNRS, Inria, I3S, France, email: {tmayer, villata}@i3s.unice.fr, elena.cabrio@unice.fr

[2] Patient Problem or Population, Intervention, Comparison or Control, and Outcome.

[3] The newly created dataset, called AbstRCT, and the annotation guidelines are available here: `https://gitlab.com/tomaye/abstrct/`

[4] The source code is available here: `https://gitlab.com/tomaye/ecai2020-transformer_based_am`

argument graphs, in which the relations connecting the retrieved argumentative components correspond to the edges. Different methods have been employed to address these tasks, from standard SVMs to NNs. AM methods have been applied to heterogeneous types of textual documents, e.g., persuasive essays [38], scientific articles [39], Wikipedia articles [4], political speeches and debates [27], and peer reviews [17]. However, only few approaches [42, 14, 25, 26] focused on automatically detecting argumentative structures from textual documents in the medical domain, such as clinical trials, clinical guidelines, and Electronic Health Records.

Few approaches consider the whole AM pipeline in different application scenarios. In particular, Stab and Gurevych [38] propose a feature-based Integer Linear Programming approach to jointly model argument component types and argumentative relations in persuasive essays. Differently from our data, essays have exactly one major claim each. The authors impose the constraint such that each claim has no more than one parent, while no constraint holds in our case. In contrast with this approach, Eger et al. [11] present neural end-to-end learning methods in AM, which do not require the hand-crafting of features or constraints, using the persuasive essays dataset. They employ TreeLSTM on dependency trees [28] to identify both components and relations between them. They decouple component classification and relation classification, but they are jointly learned, using a dependency parser to calculate the features. In our approach, we also decouple the two classification tasks, in line with the claim of [11] that decoupling component and relation classification improves the performance. Furthermore, the same work addresses component detection as a multi-class sequence tagging problem [37]. Differently from their approach, which does not scale with long texts as it relies on dependency tree distance, our approach is distance independent. In addition, whilst persuasive essay components are usually linked to components close by in the text, in our dataset links may span across the whole RCT abstract.

Recent approaches for link prediction rely on pointer networks [34] where a sequence-to-sequence model with attention takes as input argument components and returns the links between them. In these approaches, neither the boundary detection task nor the relation classification one are tackled. Another approach to link prediction relies on structured learning [12]. The authors propose a general approach employing structured multi-objective learning with residual networks, similar to approaches on structured learning on factor graphs [29]. Recently, the argument classification task was addressed with contextualized word embeddings [36]. However, differently from our approach, they assume components are given, and boundary detection is not considered. In line with their work, we experimented with the BERT [10] base model to address parts of the AM pipeline [26]. Contrary to this preliminary work, we now employ and evaluated various contextualized language models and architectures on each task to span the full AM pipeline.

3 Corpus creation

To address AM on clinical data, we rely on and extend our previous dataset [25], the only available corpus of Randomized Controlled Trial abstracts annotated with the different argument components (evidence, claims and major claims). Such corpus contains the same abstracts used in the corpus of RCT abstracts of [40], that were retrieved directly from PubMed[5] by searching for the disease name and specifying that it has to be a RCT. The first version of the corpus with coarse labels contained 919 argument components (615 evidence and 304 claims) from 159 abstracts comprising 4 different diseases (i.e., *glaucoma*, *hypertension*, *hepatitis b*, *diabetes*).

To obtain more training data, we have extracted from PubMed 500 additional abstracts following Strategy 1 in [40]. We selected *neoplasm*[6] as a topic, assuming that the abstracts would cover experiments over dysfunctions related to different parts of the human body (providing therefore a good generalization as for training instances).

Annotation was started after a training phase, where amongst others the component boundaries were topic of discussion. Gold labels were set after a reconciliation phase, during which the annotators tried to reach an agreement. While the number of annotators vary for the two annotation phases (component and relation annotation), the inter-annotator agreement (IAA) was always calculated with three annotators based on a shared subset of the data. The third annotator was participating in each training and reconciliation phase as well.

In the following, we describe the data annotation process for the argument components in the neoplasm dataset, and for the argumentative relations in the whole dataset. Table 1 reports on the statistics of the final dataset.

Dataset	#Evi	#Claim	#MajCl	#Sup	#Att
Neoplasm	2193	993	93	1763	298
Glaucoma	404	183	7	334	33
Hepatitis	80	27	5	65	1
Diabetes	72	36	11	44	8
Hypertension	59	26	9	53	2
Total	2808	1265	125	2259	342

Table 1. Statistics of the extended dataset. Showing the numbers of evidence, claims, major claims, supporting and attacking relations for each disease-based subset, respectively.

3.1 Annotation of argument components

Following the guidelines for the annotation of argument components in RCT abstracts provided in [25], two annotators with background in computational linguistics[7] carried out the annotation of the 500 abstracts on neoplasm. IAA among the annotators has been calculated on 30 abstracts, resulting in a Fleiss' kappa of 0.72 for argumentative components and 0.68 for the more fine-grained distinction between claims and evidence (meaning substantial agreement for both tasks). Example 1 shows a sample annotated abstract, where claims are written in bold, major claims are highlighted with a dashed underline, and evidence are written in italics.

Claims In the context of RCT abstracts, a *claim* is a concluding statement made by the author about the outcome of the study. It generally describes the relation of a new treatment (intervention arm) with respect to existing treatments (control arm) and is derived from the described results. *Major claims* are more a general/concluding *claim*, which is supported by more specific claims. The concluding statements do not have to occur at the end of the abstract, and may

[5] PubMed (`https://www.ncbi.nlm.nih.gov/pubmed/`) is a free search engine accessing primarily the MEDLINE database on life sciences and biomedical topics.

[6] While neoplasms can either be benign or malignant, the vast majority of articles is about malignant neoplasm (cancer). We stick with *neoplasm* as a term, since this was the MeSH term used for the PubMed query.

[7] In [15], researchers with different backgrounds (biology, computer science, argumentation pedagogy, and BioNLP) have annotated medical data for an AM task, showing to perform equally well despite their backgrounds.

also occur at the beginning of the text as an introductory *claim*, as in Example 1. Given the negligible occurrences of major claims in our dataset, we merge them with the claims for the classification task.

Evidence An *evidence* in RCT abstracts is an observation or measurement in the study, which supports or attacks another argument component, usually a *claim*. Those observations comprise side effects and the measured outcome of the intervention and control arm. They are observed facts, and therefore credible without further justifications, as this is the ground truth the argumentation is based on.

Example 1 Extracellular adenosine 5'-triphosphate (ATP) is involved in the regulation of a variety of biologic processes, including neurotransmission, muscle contraction, and liver glucose metabolism, via purinergic receptors. [In nonrandomized studies involving patients with different tumor types including non-small-cell lung cancer (NSCLC), ATP infusion appeared to inhibit loss of weight and deterioration of quality of life (QOL) and performance status]. We conducted a randomized clinical trial to evaluate the effects of ATP in patients with advanced NSCLC (stage IIIB or IV). [...] Fifty-eight patients were randomly assigned to receive either 10 intravenous 30-hour ATP infusions, with the infusions given at 2- to 4-week intervals, or no ATP. Outcome parameters were assessed every 4 weeks until 28 weeks. Between-group differences were tested for statistical significance by use of repeated-measures analysis, and reported P values are two-sided. Twenty-eight patients were allocated to receive ATP treatment and 30 received no ATP. *[Mean weight changes per 4-week period were -1.0 kg (95% confidence interval [CI]= 1.5 to -0.5) in the control group and 0.2 kg (95% CI =-0.2 to +0.6) in the ATP group (P=.002)]$_1$. [Serum albumin concentration declined by -1.2 g/L (95% CI=-2.0 to -0.4) per 4 weeks in the control group but remained stable (0.0g/L; 95% CI=-0.3 to +0.3) in the ATP group (P =.006)]$_2$. [Elbow flexor muscle strength declined by -5.5% (95% CI=-9.6% to -1.4%) per 4 weeks in the control group but remained stable (0.0%; 95% CI=-1.4% to +1.4%) in the ATP group (P=.01)]$_3$. [A similar pattern was observed for knee extensor muscles (P =.02)]$_4$. [The effects of ATP on body weight, muscle strength, and albumin concentration were especially marked in cachectic patients (P=.0002, P=.0001, and P=. 0001, respectively, for ATP versus no ATP)]$_5$.* [...] This randomized trial demonstrates that **[ATP has beneficial effects on weight, muscle strength, and QOL in patients with advanced NSCLC]$_1$**.

3.2 Annotation of argumentative relations

As a next step towards modeling the argumentative structures in the data, it is crucial to annotate the relations, i.e., directed links connecting the components. Those relations are connecting argument components to form the graph like structure of an argument. The relation is a directed link from an outgoing node (i.e., the *source*) to a target node. The nature of the relation can be supporting or attacking, meaning that the source component is justifying or undermining the target component. Links can occur only between certain components: evidence can be connected to either a claim or another evidence, whereas claims can only point to other claims (including major claims). The polarity of the relation (supporting or attacking) does not limit the possibility to what type of component a component can be connected. Theoretically, all types of relations are possible between the allowed combination pairs. Practically, some relations occur rather seldom compared to the frequency of others. The number of outgoing links from a component may exceed one. Furthermore, in rare cases, components cannot be connected at all. This can happen for major claims in the beginning of an abstract, whose function is to point out a related problem, unconnected to the outcome of the study itself.

Attack A component is attacking another one, if it is *i)* contradicting the proposition of the target component, or *ii)* undercutting its implicit assumption of significance, i.e., stating that the observed effects are not statistically significant. The latter case is shown in Example 2. Here, evidence 1 is attacked by evidence 2, challenging the generality of the prior observation.

Example 2 *[True acupuncture was associated with 0.8 fewer hot flashes per day than sham at 6 weeks,]$_1$ [but the difference did not reach statistical significance (95% CI, -0.7 to 2.4; P = .3).]$_2$*

The *partial-attack* is used when the source component is not in full contradiction, but weakening the target component by constraining its proposition. Those can be implicit statements about the significance of the study outcome, which usually occur between two claims (see Example 3). Attacks and partial-attacks are identified with a unique class for the relation classification task.

Example 3 **[SLN biopsy is an effective and well-tolerated procedure.]$_1$ [However, its safety should be confirmed by the results of larger randomized trials and meta-analyses.]$_2$**

Support All statements or observations justifying the proposition of the target component are considered as supporting the target (even if they justify only parts of the target component). In Example 1, all the evidence support claim 1.

We carried out the annotation of argumentative relations over the whole dataset of RCT abstracts, including both the first version of the dataset [25] and the newly collected abstracts on neoplasm. An expert in the medical domain (a pharmacist) validated the annotation guidelines before starting the annotation process. IAA has been calculated on 30 abstracts annotated in parallel by three annotators (the same two annotators that carried out the argument component annotation, plus one additional annotator), resulting in a Fleiss' kappa of 0.62. The annotation of the remaining abstracts was carried out by one of the above mentioned annotators.

4 The AM pipeline for clinical trials

In this section, we first describe the argument component detection and relation classification tasks, and then we report about the experimental setting to solve these tasks.

4.1 Argument Component Detection

The first step of the AM pipeline (visualized in Figure 1) is the detection of argumentative components and their boundaries. As described above, most of the AM approaches classify the type of component assuming the boundaries of argument components as given. To merge the component classification and boundary detection into one problem, we cast the component detection as sequence tagging task. Following the BIO-tagging scheme, each token should be labeled as either being at the **B**eginning, **I**nside or **O**utside of a component. As we have two component types in AM, this translates into a sequence tagging problem with five labels, i.e., *B-Claim*, *I-Claim*, *B-Evidence*, *I-Evidence* and *Outside*. To model the temporal dynamics of sequence tagging problems, usually Recurrent Neural Networks

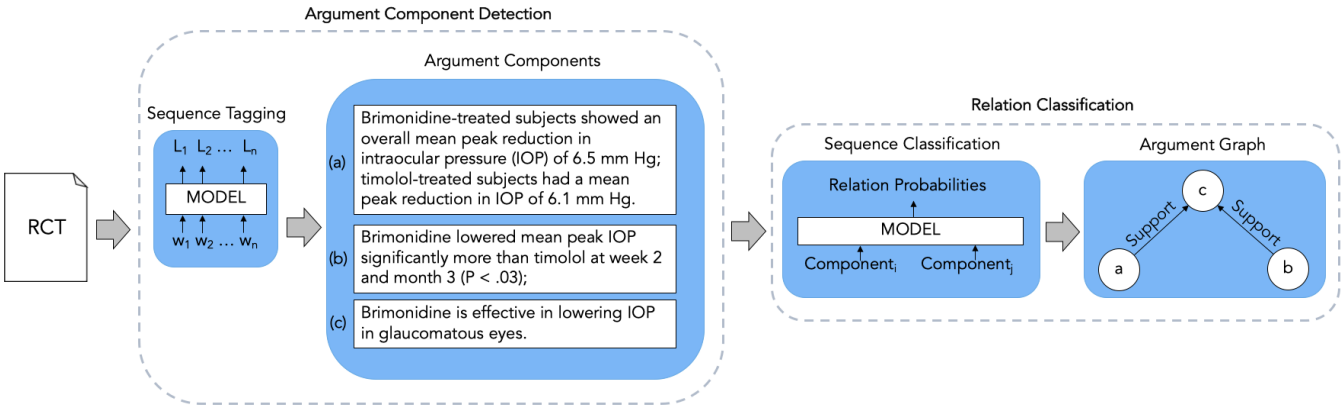

Figure 1. Illustration of the full argument mining pipeline on clinical trials.

(RNN) are used. In our experiments, we evaluate different combinations of RNNs with various types of pre-trained word representations. Each embedding method is combined with uni- or bidirectional LSTMs or GRUs with and without a CRF as a last layer. Furthermore, we are the first to do token level classification on AM by fine-tuning different transformer models.

Embeddings There are two ways to create an input word representation for sequence modelling. One way is to look up the representation from pre-trained embeddings. This static method has the advantage that one does not need to train its own embeddings. However, the vocabulary is limited, and the context of the word is not considered. State-of-the-art embeddings are generated dynamically from the context of the target word based on pre-trained language models (LM) [1, 10, 33]. In our experiments, we consider both kinds of embeddings. Furthermore, since our data is from the medical domain containing very specific terminology which might not be covered in the vocabulary of general word embeddings, we experiment with different approaches to overcome this problem.

As for the static embeddings, we employ **GloVe** [31] and **extvec** [20] embeddings, which are commonly used and are based on aggregated global word-word co-occurence statistics trained on Wikipedia and the Gigaword 5 corpus. Words are considered to be the smallest unit. Contrary to that, **fastText** [13] and byte-pair embeddings **BPEmb** [16] use subword segments to increase the capability of their vocabulary and might because of that be a better choice for a setting with unusual and specific terminology. Moving to the dynamically generated embeddings, Embeddings from Language Models (**ELMo**) [33] are generating the representation of a word by contextualizing it with the whole input sentence. They use a bidirectional LSTM to independently train a left-to-right and right-to-left character based LM. We use the ELMo model trained on PubMed to have a model which is trained on the type of data we are using. For the same reason, we use the on PubMed trained Contextualized String Embeddings (**FlairPM**) [1], another character-based language model. We compare them directly to embeddings trained on web content, Wikipedia, subtitles and news (**FlairMulti**). The third type of dynamic embedding are Bidirectional Encoder Representations from Transformers (**BERT**) [10]. The language model considers subwords and the position of the word in the sentence to give the final representation of a word.

Transformers can be used as features to an RNN, but also have the possibility to fine-tune the pre-trained model on a target dataset, which we make use of. Beside the original BERT, which is pre-trained on the BooksCorpus and English Wikipedia, there exists multiple other BERT models by now. **BioBERT** [21] is pre-trained on large-scale biomedical corpora outperforming the general BERT model in representative biomedical text mining tasks. The authors initialize the weights with the original BERT model and train on PubMed abstracts and full articles. Therefore, the vocabulary is the same as for the original BERT. Contrary to that, **SciBERT** [5] is trained from scratch with an own vocabulary. While SciBERT is trained on full papers from Semantic Scholar it also contains biomedical data, but to a smaller degree than BioBERT. We chose to use the uncased SciBERT model, meaning that we ignore the capitalization of words. As it was the case for the original BERT, the uncased model of SciBERT performs slightly better for sentence classification tasks than the cased model. Another new model, which outperforms BERT on the General Language Understanding Evaluation (GLUE) benchmark, is **RoBERTa** [23]. There, the BERT pre-training procedure is modified by exchanging static with dynamic masking, using larger byte-pair encoding and batches size, and increasing the size of the dataset.

4.2 Relation Classification

After the argument component detection, the next step is to determine which relations hold between the different components (Figure 1). We extract valid **BI** tag sequences from the previous step, which are then considered to be the argumentative components of one RCT. Those sequences are phrases and do not necessarily correspond to full sentences. The list of components then serves as input for the relation classification. As explained in Section 2, the relation classification task can be tackled with different approaches. We treat it as a sequence classification problem, where the sequence consists of a pair of two components, and the task is to learn the relation between them. For this purpose, we use self-attending transformers, since these models are dominating the benchmarks for tasks which involve classifying the status between two sentences [10]. Treating it

as a sequence classification problem gives us two options to model it: *(i)* jointly modelling the relations by classifying all possible argumentative component combinations or *(ii)* predicting possible link candidates for each entity and then classifying the relation only for plausible entity pairs. In the literature, both methods are represented. Therefore, we decided to evaluate both ways of solving the problem. We experiment with various transformer architectures and compare them with state-of-the-art AM models, i.e., the Tree-LSTM based end-to-end system from Miwa and Bansal [28] as employed by Eger et al. [11], and the multi-objective residual network of Galassi et al. [12]. For option *(i)*, we use bi-directional transformers [10], which consists of an encoder and decoder which themselves consists of multi-head self-attention layer each followed by a fully-connected dense layer. Contrary to the sequence tagging transformer, where each token of the sequence has a representation which is fed into the RNN, for sequence classification a pooled representation of the whole sequence is needed. This representation is passed into a linear layer with a softmax which decodes it into a distribution over the target classes. We treat it as a three class classification problem (*Support*, *Attack* and *NoRelation*). We refer to this type of transformer as **SentClf**. Using this architecture one component can have relations with multiple other components, since each component combination is classified independently. This is not the case in a multiple choice setting (**MultiChoice**), where possible links are predicted taking the other combinations into account and which we employ for *(ii)*. Here, each component (source) is given the list of all the other components as possible target relation candidates and the goal is to determine the most probable candidate as a target component from this list. This problem definition corresponds to the grounded common sense inference problem [43]. To model components which have no outgoing link to other components, we add the *noLink* option to the choice selection. As an encoder for phrase pairs, we evaluate various BERT models which are explained in the transformers section, just as we do for the SentClf task. With respect to the neural transformer architecture, a multiple choice setting means that each choice is represented by a vector $C_i \in R^H$, where H is the hidden size of the output of an encoder. The trainable weight is a vector $V \in R^H$ whose dot product with the choice vector C_i is the score of the choice. The probability distribution over all possible choices is given by the softmax, where n is the number of choices:

$$P_i = \frac{e^{V \cdot C_i}}{\sum_{j=1}^{n} e^{V \cdot C_j}} \quad (1)$$

The component combination with the highest score of having a link between them is then passed into a linear layer to determine which kind of relation is holding between the two components, i.e., *Attack* or *Support*. The MultiChoice model is trained jointly with two losses, i.e., one for the multiple choice task and one for for the relation classification task.

Furthermore, we experimented with linear options for link prediction, such as matrix or tensor factorization. Those methods are widely used on graph data, e.g., knowledge graphs, to discover new links between existing nodes [41]. The matrix or tensor representation of the graph data is decomposed and a model specific scoring function, which assigns a score to each triple[8], is minimized, like a loss function in neural architectures. We experiment by combining those graph-based embeddings and enriching the nodes with linguistic features/embeddings to learn hybrid graph embeddings for relations and discover new links between arguments. The tested linear models are: TuckER [3], TransE [6] and ComplEX [41]. Unfortunately, those models did not learn a meaningful relation representation. We assume this might be due to our relatively small graph data. In the literature, the smallest dataset these models have been experimented on has around 93k triples [9], whereas our dataset has less than 20k.

[8] A triple consists of a subject (source node), a predicate (labeled edge between nodes) and an object (target node).

4.3 Experimental Setup

For sequence tagging, each of the above mentioned embeddings were combined with either *(i)* a GRU, *(ii)* a GRU with a CRF, *(iii)* a LSTM, or *(iv)* a LSTM with a CRF. Additionally, the best performing static and dynamic embeddings were concatenated and evaluated as if they were one embedding. The *Flair* [1] PyTorch NLP framework version 0.4.1 was used for implementing the sequence tagging task. For BERT, we use the PyTorch implementation of huggingface[9] version 2.3. Hyper parameter tuning was done with hyperopt[10] version 0.1.2. The learning rate was selected from $\{0.05, 0.1, 0.15, 0.2\}$, RNN layers $\{1, 2\}$, hidden size $\{32, 64, 128, 256\}$, dropout $\{0.1, 0.2, 0.5\}$, and batch size from $\{8, 16, 32\}$. The RNNs were trained over 100 epochs with early stopping and SGD optimizer. For fine-tuning the BERT model, we used the uncased base model with 12 transformer blocks, a hidden size of 768, 12 attention heads, a learning rate of 2e-5 with Adam optimizer for 3 epochs. The same configuration was used for fine-tuning Sci- and BioBERT. For SciBERT, we used the uncased model with the SciBERT vocabulary. For BioBERT, we used version 1.1. For RoBERTa, we increased the number of epochs for fine-tuning to 10, as it was done in the original paper. The best learning rate was 3e-5 on our task. The number of choices for the multiple choice model was 6. Batch size was 8 with a maximum sequence length of 256 subword tokens per input example. We split our neoplasm corpus such that 350 abstracts are assigned to the train, 50 to the development, and 100 to the test set. Additionally, we use the first version of the dataset [25] to create two extra test sets, both comprising 100 abstracts. The first one includes only glaucoma, whereas the second is a mixed set with 20 abstracts of each disease in the dataset (neoplasm, glaucoma, hypertension, hepatitis and diabetes), respectively.

5 Evaluation

This section presents and discusses the empirical results of our AM pipeline for RCTs.

Sequence Tagging We show the results for the best performing RNN models and the best performing embedding combinations in Table 2. Results are given on all three test sets in micro and macro multi-class F1-score and for claim and evidence, respectively. Comparing the static word embeddings, fastText with a GRU and a CRF is the best performing combination, where extvec is only slightly worse and is usually better for evidence classification. For the dynamic embeddings coming from LMs, the ones trained on the medical domain corpus, i.e., FlairPM and ELMo, show similar performances with a macro F1-score of .68 on the neoplasm test set. They have the edge over the non-specialized LMs like BERT with .66 or FlairMulti with .63 macro F1-score. Concatenating static and dynamic embeddings does not bring a notable difference, when taking all test sets into account. Generally, evidence scores are higher than claim scores, leading to the conclusion that claims are more diverse than evidence. The

[9] https://github.com/huggingface/transformers
[10] https://github.com/hyperopt/hyperopt

		Neoplasm				Glaucoma				Mixed			
Embedding	Model	f_1	F1	C-F1	E-F1	f_1	F1	C-F1	E-F1	f_1	F1	C-F1	E-F1
GloVe	GRU+CRF	.61	.58	.50	.66	.60	.52	.36	.68	.55	.50	.36	.64
extvec	GRU+CRF	.67	.65	.58	.72	.68	.64	.57	.72	.67	.64	.57	.71
fastText(fT)	GRU+CRF	.68	.66	.61	.71	.68	.65	.60	.71	.65	.60	.52	.69
BPEmb	LSTM+CRF	.64	.60	.59	.76	.64	.60	.52	.69	.61	.57	.48	.66
ELMo	LSTM+CRF	.70	.68	.59	.76	.74	.72	.67	.77	.72	.70	.67	.74
BERT	LSTM+CRF	.69	.66	.58	.75	.70	.68	.63	.73	.68	.66	.61	.71
FlairMulti	LSTM+CRF	.66	.63	.53	.72	.58	.55	.50	.60	.52	.50	.44	.56
FlairPM	LSTM+CRF	.70	.68	.60	.75	.74	.72	.69	.75	.70	.68	.64	.72
FlairPM + extvec	GRU+CRF	.68	.65	.54	.74	.74	.72	.67	.77	.68	.66	.60	.72
FlairPM + fT	GRU+CRF	.68	.64	.53	.75	.71	.68	.62	.74	.67	.63	.56	.71
FlairPM + BERT	LSTM+CRF	.70	.69	.61	.76	.71	.70	.67	.73	.68	.67	.62	.72
BERT + fT	LSTM+CRF	.68	.65	.55	.74	.68	.66	.60	.71	.67	.65	.58	.71
ELMo + fT	LSTM+CRF	.71	.68	.59	.77	.74	.72	.69	.77	.72	.70	.65	.75
fine-tuning BERT	dense layer	.82	.60	.69	.83	.77	.55	.63	.80	.80	.57	.65	.83
fine-tuning BERT	GRU+CRF	.89	.85	.78	.90	.89	.86	.76	.89	.90	.88	.81	.91
fine-tuning BioBERT	GRU+CRF	**.90**	.84	.87	.90	**.92**	**.91**	**.93**	.91	**.92**	**.91**	**.91**	.92
fine-tuning SciBERT	GRU+CRF	**.90**	**.87**	**.88**	**.92**	.91	.89	**.93**	.91	.91	.88	.90	**.93**

Table 2. Results of the multi-class sequence tagging task are given in micro F1 (f_1) and macro F1 (F1). The binary F1 for claims are reported as C-F1 and for evidence as E-F1. Best scores in each column are marked in bold; significance was tested with a two-sided Wilcoxon signed rank test.

explanation is that, since natural language reports of measurements in clinical trials vary mostly only in the measured parameter and its values, claims can be made about almost everything. Another observation is that the performance of the models trained on neoplasm data do not significantly decrease for test sets on other disease treatments. This fact supports our choice of a more general high level disease type like neoplasm for training the models. The performance for many model combinations even increases on the glaucoma test set. The glaucoma test set comprises only a handful of different glaucoma treatments and is therefore less diversified than the neoplasm or mixed test sets. Looking at the main difference in the results, fine-tuning BERT outperforms all other model combinations, where the version with a GRU and CRF is the best performing model. Fine-tuning without any kind of sequence modelling on top of it results in worse performance. Especially with respect to the validity of BIO sequences, where disproportionately many invalid sequences are generated. This is not useful when extracting the components based on BIO-scheme. Comparing the specialized with the general models, Bio- and SciBERT show a better performance than the general BERT model, where the cased BioBERT tends to be more reliable for the out of domain test data. This is in line with the findings that the cased transformer model works better for tasks like Named Entity Recognition (NER), which is also a sequence tagging task. The difference on our data is marginal: while for NER the casing of a word is relevant, in our task it does not seem to be a sensitive information.

Relation Classification The results for relation classification are shown in Table 3. The numbers are not calculated on gold standard, but show the actual relation classification performance when the components come from the sequence tagging module of the pipeline. We used the best performing sequence tagger, i.e. the fine-tuned SciBERT with a GRU and CRF. We follow previous work on AM [32] and consider the overlap percentage of the components to determine the base if a predicted component matches the annotated component in the gold standard. Since in our data a lot of the components span over 50% or more of a sentence and the exact boundary detection is not always clear, even for human annotators, we consider a predicted and a gold standard component as matched, when at least 75% of the words overlap.

Method	Neoplasm	Glaucoma	Mixed
Tree-LSTM	.37	.44	.39
Residual network	.42	.38	.43
BERT MultiChoice	.58	.56	.55
BioBERT MultiChoice	.61	.58	.57
SciBERT MultiChoice	.63	.59	.60
BERT SentClf	.62	.53	.66
BioBERT SentClf	.64	.58	.61
SciBERT SentClf	**.68**	.62	**.69**
RoBERTa	.67	**.66**	.67

Table 3. Results of the relation classification task, given in macro F1-score.

The Tree-LSTM based end-to-end system performed the worst with a F1-score of .37. This can be explained by the positional encoding in the persuasive essay dataset being more relevant than in ours. There, components are likely to link to a neighboring component, whereas in our dataset the position of a component only partially plays a role, and therefore the distance in the dependency tree is not a meaningful feature. Furthermore, the authors specify that their system does not scale with increasing text length [11]. Especially detailed reports of measurements can make RCT abstracts quite long, such that this system becomes not applicable for this type of data.

The residual network performed better with a F1-score of .42. The main problem here is that it learns a multi-objective for link prediction, relation classification and type classification for source and target component, where the latter classification step is already covered by the sequence tagger and therefore unnecessary at this step.

Similar to sequence tagging, one can see a notable increase in performance when applying a BERT model. Comparing the specialized and general BERT model, the Bio- and SciBERT increase the performance by up to .06 F1-score. Interestingly, RoBERTa delivers com-

parable results even though it is a model trained on general data. We speculate that parts of the web crawl data which was used to train RoBERTa contain PubMed articles, since they are freely available on the web. Independently of that, RoBERTa shows more reliable results when looking at the performance on the out of domain test sets. While SciBERT as the best performing system on the in-domain test set drops .06 points on the glaucoma test set, RoBERTa stays almost the same and only drops from .67 to .66 F1-score. Looking at the difference between the MultiChoice and SentClf architectures, the SentClf delivers slightly better results, but the drawback is that this technique tends to link components to multiple components. Since most of our components have only one outgoing edge, it creates a lot of false positives, i.e., links which do not exist.

While our dataset consists of only study abstracts for practical reasons, the pipeline can be applied on full text articles as well. Alas, we cannot provide a quantitative analysis on full articles due to missing annotated data. In preliminary experiments on full articles, we have observed a notable increase of false positives in the relation classification, which is the expected consequence of an increased number of components. Furthermore, with the number of components rising in the double-digit range, the multiple-choice architecture loses its predictive power. We leave further investigations to determine the exact limit of this architecture applied on full text articles to future work.

Error Analysis Common mistakes for the sequence tagger are the invalid BIO sequences. Especially when there are multiple components in one sentence, the tagger tends to mislabel *B-* tokens as *I-* tokens. This is due to the natural imbalance between *B-* and *I-* tokens. Training the sequence tagging without the BIO scheme using only *claim* and *evidence* as labels, poses problems when multiple components are following each other in the text. They would be extracted as one single component instead. This is a common case in concluding sentences at the end of a study, which strikingly often comprise multiple claims. Further experiments could go in the direction of weighted loss functions like focal loss to overcome this problem. Notable mistakes arise for determining the exact component boundaries. Especially in the case of connectives, e.g., *however*, which have sometimes nothing but a conjunctive function, and in other cases signal a constraint of a previous statement. Another mistake is the misclassification of the description of the initial state of the participant groups as an observation of the study and therefore an evidence, e.g., *there were no significant differences in pregnancy-induced hypertension across supplement groups*. In the study abstract these descriptions occur usually relatively close to the actual result description, which means that adding information of the position in the text will not avoid this error. While only some abstracts are structured, the full study report does usually have separated sections. This structure can be exploited when analysing full reports, and in the simplest case one would analyse only the sections of interest.

Concerning link prediction, general components like *the difference was not statistically significant* are problematic, since it could be linked to most of the components/outcomes of the trial. Here, a positional distance encoding could be beneficial, since those components are usually connected to the previous component. In general, most of the errors in the MultiChoice architecture were made in the multiple choice part by predicting a wrong link and not at the stage of classifying the relation type. Interestingly, comparing the two domain adapted models, Bio- and SciBERT, there were no regular errors, which allows any conclusion about the advantages or disadvantages of one model. Looking at the confusion matrices, all tested SentClf models show a higher error rate for the *NoRelation* class. Both transformer approaches have in common the problem of dealing with negations and limitations or associating the polarity of a measurement and therefore confusing support and attack.

Example 4 **[more research about the exact components of a VR intervention and choice of outcomes to measure effectiveness is required]$_{\text{source}}$ [Conducting a pragmatic trial of effectiveness of a VR intervention among cancer survivors is both feasible and acceptable]$_{\text{target}}$**

Example 5 **[this did not translate into improved progression-free survival (PFS) or overall survival]$_{\text{source}}$ [The addition of gemcitabine to carboplatin plus paclitaxel increased treatment burden, reduced PFS time, and did not improve OS in patients with advanced epithelial ovarian cancer]$_{\text{target}}$**

Example 4 shows two claims with a limiting/attacking relation, which was wrongly classified as supporting. For Example 5, *not improving progression-free survival (PFS)* corresponds to a *reduced PFS time*, while for other factors reducing the value means it is beneficial and therefore improving some study parameter. Here, the inclusion of external expert knowledge is crucial to learn these fine nuances. The polarity of a measurement cannot be learnt from textual features alone. Especially in the medical domain, there are complex interrelationships which are not often explicitly mentioned and therefore are impossible to capture with a model trained solely on character-based input. Phrases like *increased the blood pressure by X* or *showed no symptom of Y* can connote different messages depending on the context. Future work needs to consider this challenge of incorporating external expert knowledge. While we do not think this is a problem limited to a special domain, we consider it greatly important for understanding and representing medical text.

6 Conclusion

To support clinicians in decision making or in (semi)-automatically filling evidence tables for systematic reviews in evidence-based medicine, we propose a complete argument mining pipeline for the healthcare domain. To this aim, we built a novel corpus of healthcare texts (i.e., RCT abstracts) from the MEDLINE database, which are annotated with argumentative components and relations. Indeed, we show that state-of-the-art argument mining systems are unable to satisfactorily tackle the two tasks of argument component detection and relation prediction on this kind of text, given its peculiar features (e.g., component relations spanning across the whole RCT abstract). We expect that our work will have a large impact for clinicians as it is a crucial step towards AI supported clinical deliberation at a large scale.

We employ a sequence tagging approach combining a domain specific BERT model with a GRU and CRF to identify and classify argument components. We cast the relation classification task as a multiple choice problem and compare it with recent transformers for sequence classification. In our extensive evaluation, addressed on a newly AM annotated dataset of RCTs, we investigate the use of different neural transformer architectures and pre-trained models in this pipeline, showing an improvement of the results in comparison with standard baselines and state-of-the-art AM systems.

For future work, we will annotate relations across different RCTs to allow reasoning on the resulting argument graphs and clustering of arguments about the same disease. Furthermore, we will investigate different ways to efficiently deal with medical abbreviations and incorporate a distance parameter to overcome the problem that

general components talking about limitations are linked to unrelated components far away in the text of the RCT abstract.

ACKNOWLEDGEMENTS

This work is partly funded by the French government labelled PIA program under its IDEX UCA JEDI project (ANR-15-IDEX-0001). This work has been supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002

REFERENCES

[1] Alan Akbik, Duncan Blythe, and Roland Vollgraf, 'Contextual string embeddings for sequence labeling', in *Proc. of COLING 2018*, pp. 1638–1649, (2018).

[2] Katie Atkinson, Pietro Baroni, Massimiliano Giacomin, Anthony Hunter, Henry Prakken, Chris Reed, Guillermo Ricardo Simari, Matthias Thimm, and Serena Villata, 'Towards artificial argumentation', *AI Magazine*, **38**(3), 25–36, (2017).

[3] Ivana Balazevic, Carl Allen, and Timothy Hospedales, 'TuckER: Tensor factorization for knowledge graph completion', in *Proc. of EMNLP-IJCNLP 2019*, pp. 5185–5194, (2019).

[4] Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim, 'Stance classification of context-dependent claims', in *Proc. of EACL 2017*, pp. 251–261, (2017).

[5] Iz Beltagy, Kyle Lo, and Arman Cohan, 'SciBERT: A pretrained language model for scientific text', in *Proc. of EMNLP-IJCNLP 2019*, pp. 3615–3620, (2019).

[6] Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko, 'Translating embeddings for modeling multi-relational data', in *Proc. of NIPS 2013*, pp. 2787–2795, (2013).

[7] Elena Cabrio and Serena Villata, 'Five years of argument mining: a data-driven analysis', in *Proc. of IJCAI 2018*, pp. 5427–5433, (2018).

[8] Robert Craven, Francesca Toni, Cristian Cadar, Adrian Hadad, and Matthew Williams, 'Efficient argumentation for medical decision-making', in *Proc. of KR 2012*, pp. 598–602, (2012).

[9] Tim Dettmers, Minervini Pasquale, Stenetorp Pontus, and Sebastian Riedel, 'Convolutional 2d knowledge graph embeddings', in *Proc. of AAAI 2018*, pp. 1811–1818, (February 2018).

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 'BERT: Pre-training of deep bidirectional transformers for language understanding', in *Proc. of NAACL-HLT 2019*, pp. 4171–4186, (2019).

[11] Steffen Eger, Johannes Daxenberger, and Iryna Gurevych, 'Neural end-to-end learning for computational argumentation mining', in *Proc. of ACL 2017*, pp. 11–22, (2017).

[12] Andrea Galassi, Marco Lippi, and Paolo Torroni, 'Argumentative link prediction using residual networks and multi-objective learning', in *Proc. of ArgMining 2018 workshop*, pp. 1–10, (2018).

[13] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov, 'Learning word vectors for 157 languages', in *Proc. of LREC 2018*, pp. 3483–3487, (2018).

[14] Nancy Green, 'Argumentation for scientific claims in a biomedical research article', in *Proc. of ArgNLP 2014 workshop*, (2014).

[15] Nancy Green, 'Annotating evidence-based argumentation in biomedical text', *IEEE BIBM 2015*, 922–929, (2015).

[16] Benjamin Heinzerling and Michael Strube, 'Bpemb: Tokenization-free pre-trained subword embeddings in 275 languages', in *Proc. of LREC 2018*, pp. 2989–2993, (2018).

[17] Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang, 'Argument mining for understanding peer reviews', in *Proc. of NAACL-HLT 2019*, p. 2131–2137, (2019).

[18] Anthony Hunter and Matthew Williams, 'Aggregating evidence about the positive and negative effects of treatments', *Artificial Intelligence in Medicine*, **56**(3), 173–190, (2012).

[19] Di Jin and Peter Szolovits, 'PICO element detection in medical text via long short-term memory neural networks', in *Proc. of BioNLP 2018 workshop*, pp. 67–75, (2018).

[20] Alexandros Komninos and Suresh Manandhar, 'Dependency based embeddings for sentence classification tasks', in *Proc. of NAACL-HLT 2016*, pp. 1490–1500, (2016).

[21] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang, 'BioBERT: a pre-trained biomedical language representation model for biomedical text mining', *Bioinformatics*, (2019).

[22] Marco Lippi and Paolo Torroni, 'Argumentation mining: State of the art and emerging trends', *ACM Trans. Internet Techn.*, **16**(2), 10, (2016).

[23] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, 'Roberta: A robustly optimized BERT pretraining approach', *CoRR*, **abs/1907.11692**, (2019).

[24] Luca Longo and Lucy Hederman, 'Argumentation theory for decision support in health-care: A comparison with machine learning', in *Proc. of BHI 2013*, pp. 168–180, (2013).

[25] Tobias Mayer, Elena Cabrio, Marco Lippi, Paolo Torroni, and Serena Villata, 'Argument mining on clinical trials', in *Proc. of COMMA 2018*, pp. 137–148, (2018).

[26] Tobias Mayer, Elena Cabrio, and Serena Villata, 'ACTA a tool for argumentative clinical trial analysis', in *Proc. of IJCAI 2019*, pp. 6551–6553, (2019).

[27] Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata, 'Never retreat, never retract: Argumentation analysis for political speeches', in *Proc. of AAAI 2018*, pp. 4889–4896, (2018).

[28] Makoto Miwa and Mohit Bansal, 'End-to-end relation extraction using lstms on sequences and tree structures', in *Proc. of ACL 2016*, pp. 1105–1116, (2016).

[29] Vlad Niculae, Joonsuk Park, and Claire Cardie, 'Argument mining with structured SVMs and RNNs', in *Proc. of ACL 2017*, pp. 985–995, (2017).

[30] Andreas Peldszus and Manfred Stede, 'From argument diagrams to argumentation mining in texts: A survey', *Int. J. Cogn. Inform. Nat. Intell.*, **7**(1), 1–31, (2013).

[31] Jeffrey Pennington, Richard Socher, and Christopher D. Manning, 'Glove: Global vectors for word representation', in *Proc. of EMNLP 2014*, pp. 1532–1543, (2014).

[32] Isaac Persing and Vincent Ng, 'End-to-end argumentation mining in student essays', in *Proc. of NAACL-HLT 2016*, pp. 1384–1394, (2016).

[33] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer, 'Deep contextualized word representations', in *Proc. of NAACL-HLT 2018*, pp. 2227–2237, (2018).

[34] Peter Potash, Alexey Romanov, and Anna Rumshisky, 'Here's my point: Joint pointer architecture for argument mining', in *Proc. of EMNLP 2017*, pp. 1364–1373, (2017).

[35] Malik Al Qassas, Daniela Fogli, Massimiliano Giacomin, and Giovanni Guida, 'Analysis of clinical discussions based on argumentation schemes', *Procedia Computer Science*, **64**, 282–289, (2015).

[36] Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych, 'Classification and clustering of arguments with contextualized word embeddings', in *Proc. of ACL 2019*, pp. 567–578, (2019).

[37] Anders Søgaard and Yoav Goldberg, 'Deep multi-task learning with low level tasks supervised at lower layers', in *Proc. of ACL 2016*, pp. 231–235, (2016).

[38] Christian Stab and Iryna Gurevych, 'Parsing argumentation structures in persuasive essays', *Comput. Linguist.*, **43**(3), 619–659, (2017).

[39] Simone Teufel, Advaith Siddharthan, and Colin Batchelor, 'Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics', in *Proc. of EMNLP 2009*, pp. 1493–1502, (2009).

[40] Antonio Trenta, Anthony Hunter, and Sebastian Riedel, 'Extraction of evidence tables from abstracts of randomized clinical trials using a maximum entropy classifier and global constraints', *CoRR*, **abs/1509.05209**, (2015).

[41] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard, 'Complex embeddings for simple link prediction', in *Proc. of ICML 2016*, pp. 2071–2080, (2016).

[42] Jure Zabkar, Martin Mozina, Jerneja Videcnik, and Ivan Bratko, 'Argument based machine learning in a medical domain', in *Proc. of COMMA 2006*, pp. 59–70, (2006).

[43] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi, 'SWAG: A large-scale adversarial dataset for grounded commonsense inference', in *Proc. of EMNLP 2018*, pp. 93–104, (2018).